Emerging methods for conceptual modelling in neuroimaging

Akama, H., & Murphy, B. (2016). Emerging methods for conceptual modelling in neuroimaging. *Behaviormetrika*. https://doi.org/10.1007/s41237-016-0009-1

Published in:
Behaviormetrika

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
The Behaviormetric Society 2016
The final publication is available at Springer via link.springer.com/article/10.1007/s41237-016-0009-1

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

Download date:04. May. 2023

Review Paper

Behaviormetrika
DOI 10.1007/s41237-016-0009-1

Emerging Methods for Conceptual Modelling in Neuroimaging

Hiroyuki Akama[1], Brian Murphy[2]

1: Institute of Liberal Arts/School of Life Science and Technology , Tokyo Institute of Technology, Tokyo, 152-8552, Japan
2: Knowledge & Data Engineering, EEECS, Queen's University Belfast, BT9 5BN, Northern Ireland

Abstract

Some open theoretical questions are addressed on how the mind and brain represent and process concepts, particularly as they are instantiated in particular human languages. Recordings of neuroimaging data should provide a suitable empirical basis for investigating this topic, but the complexity and variety of language demands appropriate data-driven approaches. In this review we argue for a particular suite of methodologies, based on multivariate classification techniques which have proven to be powerful tools for distinguishing neural and cognitive states in fMRI. A combination of larger scale neuroimaging studies are introduced with different monolingual and bilingual populations, and hybrid computational analyses that use encoded implementations of existing theories of conceptual organisation to probe that data. We develop a suite of methodologies that holds the promise of being able to holistically elicit, record and model neural processing during language comprehension and production.

Keywords: fMRI, MVPA, Machine learning, Embodiment, **Cross-subject analysis**

Correspondence: Dr. Hiroyuki Akama, Institute of Liberal Arts/ School of Life Science and Technology, Tokyo Institute of Technology, W9-10, 2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan, Mail: akama.h.aa@m.titech.ac.jp Tel & Fax: 81-3-5734-3254.

Conflicts of interest: none.

I. Introduction

The question of how the human mind stores and represents the meanings of concepts, and of the words we use to describe them, remains something of an enigma. Historically, approaches range from the metaphysical, such as Platonic models that assume an inherent and universal organisation of reality (e.g. Wilkins, 1668), to the highly empirical, such as purely data-driven approaches that derive conceptual spaces from arbitrary web-text (Lund & Burgess 1996; Landauer & Dumais 1997; Griffiths et al 2007; Carlson et al 2010; Murphy et al 2012a). Contemporary debate revolves around questions such as taxonomic membership, whether concepts are represented by concrete exemplars or abstract prototypes, embodied/symbolic encodings, and whether property representations are holistic, or composed of distributed activations or atomic features (see Murphy, 2004 for a comprehensive overview).

In this review we will argue for a particular suite of methodologies to investigate these questions: that is a combination of larger scale neuroimaging studies with different monolingual and bilingual populations, and hybrid computational analyses that use encoded implementations of existing theories of conceptual organisation to probe the neural data. In particular we present experiments that look at how to deal with variation seen in the neural signatures associated with broad semantic classes (animals and tools) from recording to recording, depending on the presentation modality of the stimulus (written vs spoken words), and the language used (in Japanese/Chinese, and Chinese/Korean bilinguals). We present some methodological innovations that aid this aim, and finish with a discussion of how a comprehensive investigation of these broader questions may proceed.

II. Language and Conceptual Organisation

Two questions in particular are of current interest to the cognitive science and neuroscience communities. One is whether meaning is encoded entirely with **abstract symbols**, or entirely in terms of direct **embodied experience** (see section below), with less extreme positions taking elements of both approaches. Another is the extent to which conceptual organisation is determined by a shared genetic endowment and so **universal**; or shaped by different life experience, and so **culturally specific**.

Embodiment Theory asserts that the meaning of words is grounded and situated in our experience world in which we live, rather than being encoded in *amodal* abstract symbols (see Meteyard et al 2012 for a critical review). The neuro-physiological significance of this theory consists in a simulation semantics, which claims that conceptual processing of a word is performed by activating the perceptuo-motor system in the brain to virtually experience a context of its referent. Support for embodiment theory can be seen in some neuroimaging data (e.g. Pulvermüller 2005). For instance, concepts which involve a motor action as part of their core semantics, such as hand-tools, may elicit activity in the motor areas activated during their use (see Figure 1).

Recordings of brain activity have a unique advantage when it comes to investigating the mind, since they are a real-space and real-time recording of mental processes. However brain data recording techniques are limited, in that none of them comprehensively capture the activity of the whole brain in action, in effect aggregating and blurring the activity of individual neurons over time and over space. Existing

methods are vulnerable to various kinds of environmental and instrumental noise, and involve a trade-off between good spatial resolution (e.g. fMRI) or good temporal resolution (e.g. EEG). No current or projected methods approach a clean and comprehensive recording of the whole brain in action, which is comprised by the simultaneous activations of tens of billions individual neurons. Furthermore, brain imaging data can be expensive and cumbersome to collect, meaning that the datasets used in cognitive neuroscience research are typically very small (e.g. 30-60mins of activity from each of 10-20 participants). This has proven sufficient to answer many basic questions of cognitive function, typically using an artificial and targetted task that homes in on a single aspect of psychology at a time, using manipulated and carefully controlled sets of stimuli. A working assumption of such studies is often that human participants are homogeneous, so that cognitive processing is assumed to be basically uniform in kind across genders, age groups, cultures and other groupings.

Such an approach may not be appropriate for complex higher-level cognitive systems (Newell 1973), and may be especially ill-suited to investigations of language. Human languages are based in cultural experience, and there is some evidence (though still controversial, see e.g. Nevins et al 2009) that nurture partially determines conceptual organisation of constructs including number, time and colour terms (e.g. Regier et al 2007; Frank et al 2008). Further, mental vocabularies consist of tens of thousands of lexemes, expressing many thousands of concepts[1]. Even the core syntax of well-studied languages such as English have defied comprehensive treatment in theories that take a reductionist approach (e.g. Chomsky 1981).

III. Data-Driven Analysis of Neuroimaging Data

Important progress has been made in cognitive neuroscience by applying partially data-driven computational approaches to recordings of brain activity. Machine learning methods, often termed multivariate pattern analysis (MVPA), particularly in the fMRI community, have several advantages over conventional contrastive analyses, summarized below.

Increased Sensitivity and Adaptivity: data-driven analyses may identify unanticipated patterns in the data, rather than those hypothesised by those who carry out experiments.

Individual Stimulus Analysis: increased sensitivity can make analysis of the neural signatures of individual stimuli possible, e.g. single words, or images.

Individual Participant Analysis: the same sensitivity enables analysis of data from single human participants.

Within-Study Replication: when cross-validated machine learning is used, there is a degree of replication, in that the patterns of activity identified during training are validated on unseen data.

---

1 For instance, WordNet, a large scale computational lexicon of contemporary English, contains over 200 thousand word-sense entries. See http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html.

Multi-factor Modelling of Stimuli: rich meaningful stimuli (even stories and films) can be comprehensively modelled, as an alternative to attempts at balance of confounds.

Haxby et al (2001) was perhaps the first successful machine learning analysis of fMRI data. Using a simple two-fold data partition method, the study took odd and even runs of the presentation of visual object stimuli in several categories (e.g. shoes, houses, chairs). A transparent pattern identification method found reasonably stable **neural signatures –** that is distributed patterns of activity associated with a particular stimulus or psychological state. After learning these neural signatures, the visual category of unseen trials of fMRI data could be discriminated using a simple nearest-neighbour classifier.

A further advance was represented by Mitchell et al (2008), which succeeded in modelling contentful stimuli (a concept, represented by a line-drawing and its word-label) in terms of dimensions of meaning. A simple vector space feature model described the semantics of each concept, and a linear regression model was used to decompose whole-brain activity to learn neural signatures for these individual dimensions. So, for example, that study was able to identify the distributed pattern of activity typical of objects which were edible (fruit, vegetables and certain animals). In this way, the study demonstrated the ability to **extrapolate** beyond the limited set of stimuli that can be presented in a single neuroimaging session. Specifically, using a leave-two-out cross-validation partition, it could discriminate which of two concepts (e.g. *horse* and *car*) a participant was thinking about, even though neither of those concepts had been present in the training data.

One consequence of the increased sensitivity of data-driven methods is that they adapt to idiosyncrasies in particular participants (e.g. of anatomy, or timing in processing), or recording sessions (e.g. environmental noise). This can make generalisations over multiple participants and session more challenging (conventional analyses typically correct for this with strategies involving temporal and spatial averaging and smoothing). Haxby et al (2011) introduced a new method for cross-participant normalisation, based on richer, more ecologically valid stimulus: popular film. The hyperalignment method also used regression-based techniques to identify a set of voxels in one participant whose collective activity corresponded to that of the participant group. This was done on the assumption that popular film (in this case the Steven Spielberg feature, *Indiana Jones and the Temple of Doom*), drives neural activity such that whole-brain activations should be **synchronised** across individuals. Hyperalignment takes advantage of this to provide a **functional mapping** between the individual participants and the group at large, enabling much more effective cross-group learning of neural signatures.

IV. Variation in Neural Signatures across Participants and Experiments

In principle, one would hope that the neural signatures observed for a particular concept would exhibit a large degree of stability in repeated experimental measurements.

Whether conceptual representations are shared across individuals, language populations and communicative tasks is a central question for embodiment theories and for the question of language universality/relativity. At the same time, there are reasons to believe there is some degree of commonality. Neuroimaging and lesions studies confirm that at a macro-scale, cortical localisation of principal cognitive functions is largely shared across individuals (at least when limited to certain homogeneous populations, e.g.

those with shared handedness, developmental stage, and health status). Arguments of cognitive and physiological economy suggest that the brain would develop in such a way as to avoid redundantly replicating conceptual information associated with different modalities such as vision, speech, and writing (though competing pressures of speed of processing, and robustness to cognitive degradation would encourage some degree of redundancy). Further, individuals switch easily between tasks and modalities, bilingual individuals switch languages, and groups of people communicate, all suggesting a partially shared semantic base.

Even if conceptual organisation was identical in all situations and participants, more prosaic differences would obscure them. Cortical anatomy varies considerably across individuals, in terms of shape, size, and local cortical folding, and this is only partially compensated for with image registration methods. And the temporal profile of activations may be modulated by the task, physiological arousal, and attentional state.

In the experiments presented here (Akama et al 2012; Akama et al 2014), we demonstrate some of these issues. In Figure 2, we show the binary average decoding accuracy for a mammal/tool classification task (mammal/tool, using ANOVA feature selection from full-brain data, and regularised logistic regression – see Methods section for full details) from a series of within-participant analyses of Experiment I. Chance performance is 50%, and accuracies above 55.8% were significant (at $p<0.05$, binomial test over independent trials, chance 50%, n=240). Note that experimental results are presented throughout this review, and our methods are described at the end.

As can be seen there is some variation already between different individual participants (P1-P5). Task effects are seen in two ways. Overall, when cross-validated training and testing data come from the same session, the auditory condition (audio-audio: spoken word stimulus) seems more easily decodable than the orthographic one (ortho-ortho: written word stimulus). And when we test and train across two tasks from the same participant (training using the auditory-condition session and testing with the orthographic-condition session, or vice versa), decoding performance is further degraded, suggesting systematic differences in the neural signatures observed.

This variation has numerous possible causes, but differences in spatial and temporal localisation appear to contribute. Figure 3 performs the same analysis as in Figure 2, but using only a single fMRI trial volume at a time as input to the classifier (corresponding to a single second of data recorded in the first 20 seconds after stimulus presentation). Interestingly, the temporal profile of classification accuracy, which corresponds to the availability of discriminative information in the recorded signals, is isomorphic with the default hemodynamic function used in packages such as SPM. However there is considerable variation across participants (dotted lines), and the auditory-stimulus task seems to elicit a stronger and longer lasting neural signature, relative to the orthographic-stimulus task.

Informative voxels are also widely spatially distributed. In Figure 4 we see the proportion of voxels that were chosen by an ANOVA feature ranking within selected functional/anatomical regions according to the AAL atlas (Tzourio-Mazoyer et al, 2002), averaged over all ten recorded sessions in Experiment I. When we differentiate by task modality, these spatial concentrations vary also, as Figure 5 illustrates. While there is some commonality, the auditory-stimulus data has an stronger concentration in parietal-temporal areas, and the orthographic-stimulus data in occipital areas.

If we attempt to train and test classifiers across different participants these problems of transferability of learning become more pronounced. In Experiments IIa and IIb (Akama et al 2014), we recorded 28 sessions in total, 2 each from 14 participants. Single session cross-validated classifications, and within-participant cross-session classifications gave similar results to those seen in Figure 2. Figure 6 shows cross-session classification accuracies for all pairs of sessions. Cross-participant accuracies are clearly lower than within-participant analyses. As an aside, a strong language-specific effect was not observed – i.e., it is not the case that training and testing within the subgroups (early Korean-Chinese bilinguals, and late Chinese-Japanese bilinguals) was substantially easier, confirming that the patterns used by the algorithm are not based in language-specific word-forms, but rather in a some cross-language representation, presumably semantics.

Figure 7 shows the corresponding classification performance if we train on a whole group of participant sessions, and test a single left-out participant. The increase in data quantity, and generality of the neural patterns observed, improves generalisation across individuals. But it does not completely eliminate the cross-participant learning penalty. Figure 8 summarizes the results so far from Experiments IIa and IIb so far: the right-most bar plots show the average classification accuracy for within-subject analyses (>90%), which drops dramatically for cross-participant learning (~65%, corresponding to the results in Figure 6). Taking advantage of larger groups of participants helps only somewhat, yielding ~75% accuracy.

V. Multivariate Solutions to Data Variability

As we have seen up to this point, neural signature learning can be most successful when performed on data that comes from a single recording session (in a cross-validated fashion, avoiding double-dipping; see Kriegeskorte et al, 2009). Even if the data comes from a single participant, and a very similar task, there is a considerable decoding performance penalty, in the 5-10% range in the experiments described here (Figures 2 and 8). When we train across pairs of participants the penalty increases dramatically, and is only partly offset when we expand and generalise the training data to draw from groups of participant sessions.

One solution to this issue to that has already been introduced is the hyperalignment method from Haxby et al (2011). This uses a parallel time-courses of fMRI data from two or more participants/sessions, driven by an engaging task such as movie watching. Linear regression is used identify subgroups of voxels which are collectively correlated, resulting in a cross-session mapping. In effect this makes no assumptions about spatial correspondence across sessions, but rather demands temporal alignment of the activation profile.

An alternative approach is to abstract away from the temporal and spatial particulars of a particular recording, only representing the mutual similarity relationships among neural signatures (Kriegeskorte et al 2008a; Raizada & Connolly 2012), in a similar way to the word-spaces mentioned at the beginning of this article. With such a method we might discover for example that similarity measures between “cat” and “dog” are higher, than to “cow”, and that this pattern is shared across sessions, participants, tasks, recording modalities, and even species (Kriegeskorte et al 2008b).

In Akama et al (2014) we introduce another method for cross-session and cross-group learning of neural signatures. Crucially this method requires the training data to

include some data from the target session, avoiding double-dipping by keeping a separate partition of the target data held-out for testing. While this may not be suited to many real world applications (e.g. Brain-Computer-Interaction) for basic cognitive neuroscience research it does not present problems. And as Figure 8 demonstrates, both variants of the approach, JRFS (joint ranking feature selection) and DJFS (disjoint feature selection) come close to eliminating the cross-participant penalty, if not the cross-session penalty.

VI. Closing Discussion

In this review we point to some open theoretical questions on how the mind and brain represent and process concepts, particularly as they are instantiated in particular human languages. Recordings of neuroimaging data should provide a suitable empirical basis for investigating this topic, but the complexity and variety of language demands appropriate data-driven approaches. Conventional contrastive analyses of neuroimaging data successfully generalise over groups of participants, but usually at the cost of radical simplification of the theoretical questions (cf. the concerns of Newell, 1973).

Multivariate approaches can provide a solution to the limitations of contrastive analyses, in that they can be tailored to the granularity of the question at hand. Furthermore they can integrate existing models of language (e.g. lexicon, syntax) to comprehensively test a complete instantiation of a theory. This can have the additional benefit of increasing the generality of the results obtained – a case in point is the data-driven decomposition of fMRI brain activity in Mitchell et al (2008), which used a web-corpus derived semantic space to identify neural signatures for individual components of meaning, and so make the models capable of accounting for concepts/stimuli that have not been encountered during training (so called “zero-shot” learning). While multivariate approaches have a potential drawback in their ability to generalise across sessions, participants and tasks[2], proposals have also been made by ourselves and others (Kriegeskorte et al 2008a; Haxby et al 2011; Akama et al 2014) to deal with this effectively. Over the past few years a significant progress is being made in this field, and it is worthwhile to emphasize that the involvement of widespread cortical areas were elucidated by using Multivariate approaches to multilingual processing (Lei et al., 2014; Pliatsikas et al, 2016).

Looking to the future, the sophistication of behavioural paradigms and multivariate analyses continues to advance. There is an increased use of “naturalistic” and “ecologically valid” tasks, which try to engage participants and elicit brain activity in a more realistic way, e.g. through the use of film (Hasson et al 2008) or stories (Brennan et al 2012). These paradigms also enable more realistic and comprehensive models of language processing: for instance Wehbe et al (2014a) combines data-driven learning of temporal profiles in fMRI, cross-subject adaptive learning, and multi-strata modelling of word, sentence and narrative level representations and their interaction. And more

---

2

It should be remembered that conventional contrastive analyses generalise at the cost of fidelity to the original data. They usually use a variety of spatial smoothing (using Gaussian filters), temporal smoothing (through use of a single assumed hemodynamic model) and averaging, which blurs real and partly systematic patterns in the data.

explicit computational models of how word meanings combine to arrive at phrase and sentence meaning have been applied to fMRI data (Fyshe et al 2015; Anderson et al 2016). Wehbe et al (2014b) applies non-linear effects and deep learning to a similar paradigm. Approaches such as these are now starting to be applied to electrophysiological data also (Dmochowski et al 2012; Wehbe et al 2014b) which enables analyses at the temporal scale of natural language (in spontaneous speech we typically produce 3-4 words a second). And finally the modelling of the idiosyncratic semantic spaces of individual participants (Charest et al 2014) demonstrates that the subtle differences in concept meaning seen in different languages may be open to analysis, despite cross-language similarities in neural representations (Buchweitz et al 2012; Zinszer et al 2016) These tools continue to provide further evidence on the amodal (Fairhall et al., 2013; Liuzzi et al., 2015) or embodied (Fernandino et al., 2015, 2016; Anderson et al., 2016) encoding of meaning of the brain.

Together, we believe that this suite of methodologies holds the promise of being able to holistically elicit, record and model neural processing during language comprehension and production.

Appendix

Methods

These experiments partially replicate the paradigm developed in Mitchell et al (2008), using fast whole-brain fMRI imaging, and a slow event-related property generation task. Details of the variations in the language and modality specific to each experiment are described below. In Experiment I, native Japanese speakers performed a monolingual task. In Experiments IIa and IIb, Korean-Chinese and Chinese-Japanese bilinguals performed a language switching task.

Task and Stimuli

The same slow event-related design was used in all experiments. Each session had 6 repeated runs for a total of 240 trials. In each trial, a concept was presented for 3 seconds followed by a fixation cross for 7 seconds. There were six additional presentations of a fixation cross of 40 seconds each, distributed just after each run, to establish a signal baseline for subsequent analysis. During the stimuli presentation, participants were asked to do a silent property generation task by thinking of appropriate features of the corresponding concept. During the fixation cross participants were asked to fixate their eyes on the cross silently and no response was required.

We employed a stimulus set representing concepts in the two categories of land-mammals and work tools based on an earlier EEG experiment (Murphy et al 2011), with 20 concepts in each of the two categories.[3] For trials that used image stimuli, 40 contrast-normalized grey-scale photographs were used.

In Experiment I (Akama et al 2012) five native Japanese speakers (ages 39-53 years) participated, each performing two separate scanning sessions on two different days separated by at least 1 week. The sessions alternated the language stimulus modality: first viewing pictures while listening to the spoken word describing the represented object (the auditory condition), and next viewing pictures with an accompanying caption (the orthographic condition). They were asked to silently enumerate properties that are characteristic of the presented concept.

In Experiment IIa and IIb (Akama et al 2014; see also Lei et al 2014) bilinguals performed the same property generation task, with an added language switch condition, which alternated between each of two sessions. All 14 participants were native speakers of Mandarin Chinese (aged 22 to 28). In experiment IIa, seven early Korean-Chinese bilinguals performed both Korean-Chinese and Chinese-Korean switch conditions (see Figure 9), and in experiment IIb seven late Chinese-Japanese bilinguals performed the corresponding switch conditions involving those two languages. The experimental stimuli used in both variants are summarized in Figure 10.

Data Acquisition

Functional MRI scans were performed with a 3.0-T General Electric Signa scanner at Tokyo Institute of Technology, Japan with an 8-channel high- resolution head coil. Functional scanning was performed using an echo planar imaging sequence with a 1000-ms repetition time (TR), 30-ms echo time (TE), and 60- degree flip angle (FA). Each volume consisted of 15* 6- mm- thick slices with an interslice gap of 1 mm; FOV: , 20×*20 cm; size of acquisition matrix, 64×*64; NEX:, 1.00. The parameter values of the anatomical scans were TR=7.284 ms, TE=2.892 ms, FA=11 degrees, Band Width=31.25 kHz, and voxel size=1 mm isotropic. Following the settings used by Mitchell et al. (personal communication), we set oblique slices in the sagittal view with a tilt of -20 to -30 degrees such that the most inferior slice was passed above the eyes and through the posterior cerebellum.

Preprocessing and Contrastive Group Analysis with GLM

fMRI initial data processing was performed with Statistical Parametric Mapping software (SPM8, Wellcome Department of Cognitive Neurology, London, UK). The data were motion corrected, co-registered to the anatomical images, segmented to identify grey matter, and normalized into standard Montreal Neurological Institute (MNI) space at

---

3

The concepts used were the following: **Mammals**: *anteater, armadillo, beaver, camel, deer, elephant, fox, giraffe, gorilla, hare, hedgehog, hippopotamus, kangaroo, koala, mole, monkey, panda, rhinoceros skunk, zebra.*

Tools: *Allen key, axe, chainsaw, craft knife, file, hammer, nail, paint roller, plaster trowel, pliers, plunger, power drill, rake, saw, scraper, scissors, screw sickle, spanner, tape measure.*

The corresponding Chinese and Korean terms are detailed in Akama et al (2014).

a resliced voxel size of 3x3x6 mm, coarse rendering for temporal resolution acuity. A General Linear Model (Friston et al., 1994) was used for conventional contrastive analyses (e.g. task > rest; mammal > tool – see Figure 1 for an example). Both single-session analyses, and random-effects group analyses were performed. Some figures were produced using the XjView toolbox[4] in addition to SPM8.

Machine Learning Analyses

The PyMVPA 2.0 package (Hanke et al 2009)[5] was used for machine learning (ie multivariate pattern) analyses. The realigned, co-registered, segmented, and grey-matter-masked (but not smoothed) images of each participant in each session were used. The time-course of each voxel was *z*-normalized and detrended, and unless otherwise noted, trialwise estimates were calculated with a boxcar average. In all analyses datasets were split into a training set and an evaluation set using a leave-one-out six-fold cross-validation. The classifier used was a penalised logistic regression (PLR) using *L2*-norm regularization. The regularization term deals with both high dimensionality and redundancy in data, and its logistic function is optimized to fit discrete data categories. In **decoding** analyses a univariate feature-selection (ANOVA, which is monotonic with the t-statistic) preceded training, and performance was evaluated as the mean percentage correct classification of the semantic category (mammal or tool) of left-out data trials. In **sensitivity analyses** either the raw ANOVA feature selection ranking was used, or mammal/tool classification accuracy in a search-light sphere (Kriegeskorte et al, 2006).

References

Akama, H., Murphy, B., Li, N., Shimizu, Y., & Poesio, M. (2012). Decoding semantics across fMRI sessions with different stimulus modalities: a practical MVPA study. Frontiers in neuroinformatics, 6.

Akama, H., Murphy, B., Lei, M. M., & Poesio, M. (2014). Cross-participant modelling based on joint or disjoint feature selection: an fMRI conceptual decoding study. In Applied Informatics (Vol. 1, No. 1, p. 1). Springer.

Anderson, A. J, Binder, J, Fernandino, L., Humphries, C., Conant, L., Aguilar, M., Wang, X., Doko, D. and Raizada, R.D.S. (2016). Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation. Cerebral Cortex, Advance Online Publication. DOI: 10.1093/cercor/bhw240.

Brennan, J., Nir, Y., Hasson, U., Malach, R., Heeger, D. J., & Pylkkänen, L. (2012).

4 http://www.alivelearn.net/xjview8/

5 Python package developed to apply machine-learning to human neurological recordings http://www.pymvpa.org/.

Syntactic structure building in the anterior temporal lobe during natural story listening. Brain and language, 120(2), 163-173.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E. R., & Mitchell, T. M. (2010). Toward an Architecture for Never-Ending Language Learning. Artificial Intelligence, 2(4), 3–3.

Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D., & Kriegeskorte, N. (2014). Unique semantic space in the brain of each beholder predicts perceived similarity. Proceedings of the National Academy of Sciences of the United States of America, 111(40), 14565–70.

Chomsky, N. (1981). Lectures on government and binding: The Pisa lectures. Walter de Gruyter.

Dmochowski, J. P., Sajda, P., Dias, J., & Parra, L. C. (2012). Correlated components of ongoing EEG point to emotionally laden attention–a possible marker of engagement?. Frontiers in human neuroscience, 6.

Fairhall, S. L. and Caramazza, A. (2013). Brain regions that represent amodal conceptual knowledge. Journal of Neuroscience, 33(25):10552–10558.

Fernandino, L., Humphries, C. J., Seidenberg, M. S., Gross, W. L., Conant, L. L., & Binder, J. R. (2015). Predicting brain activation patterns associated with individual lexical concepts based on five sensory-motor attributes, Neuropsychologia, 76:17-26

Fernandino, L., Binder, J. R., Desai, R. H., Pendl, S. L., Humphries, C. J., Gross, W. L., Conant, L. L., & Seidenberg, M. S. (2016). Concept Representation Reflects Multimodal Abstraction: A Framework for Embodied Semantics, Cerebral Cortex, 26(5), 2018-34.

Frank, M. C., Everett, D. L., Fedorenko, E., & Gibson, E. (2008). Number as a cognitive technology: Evidence from Pirahã language and cognition. Cognition, 108(3), 819-824.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D., & Frackowiak, R. S. (1994). Statistical parametric maps in functional imaging: a general linear approach. Human brain mapping, 2(4), 189-210.

Fyshe, A., Wehbe, L., Talukdar, P., Murphy, B., & Mitchell, T. (2015). A Compositional and Interpretable Semantic Space. In Proceedings of NAACL 2015 (pp. 32–41).

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. Psychological Review, 114(2), 211–244.

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V., & Pollmann, S. (2009). PyMVPA: A python toolbox for multivariate pattern analysis of fMRI data. Neuroinformatics, 7(1), 37-53.

Hasson, U., Landesman, O., Knappmeyer, B., Vallines, I., Rubin, N., & Heeger, D. J. (2008). Neurocinematics: The neuroscience of film. Projections, 2(1), 1-26.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science, 293(5539), 2425-2430.

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., Hanke, M., & Ramadge, P. J. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron, 72(2), 404-416.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. Proceedings of the National Academy of Sciences of the United States of America, 103(10), 3863-3868.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008a). Representational similarity analysis–connecting the branches of systems neuroscience. Frontiers in systems neuroscience, 2.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. A. (2008b). Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron, 60(6), 1126-1141.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. Nature neuroscience, 12(5), 535-540.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2), 211–240.

Lei, M., Akama, H., & Murphy, B. (2014). Neural basis of language switching in the brain: fMRI evidence from Korean–Chinese early bilinguals. Brain and language, 138, 12-18.

Liuzzi, A. G., Bruffaerts, R., Dupont, P., Adamczuk, K., Peeters, R., De Deyne, S., Storms, G., & Vandenberghe, R. (2015). Left perirhinal cortex codes for similarity in meaning between written words: Comparison with auditory word input. Neuropsychologia, 76, 4-16.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instruments, and Computers, 28, 203–208.

Meteyard, L., Cuadrado, S. R., Bahrami, B., & Vigliocco, G. (2012). Coming of age: a review of embodiment and the neuroscience of semantics. Cortex, 48(7), 788-804.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K. M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. science, 320(5880), 1191-1195.

Murphy, B., Poesio, M., Bovolo, F., Bruzzone, L., Dalponte, M., & Lakany, H. (2011). EEG decoding of semantic category reveals distributed representations for single concepts. Brain and language, 117(1), 12-22.

Murphy, B., Talukdar, P., & Mitchell, T. (2012a). Selecting corpus-semantic models for neurolinguistic decoding. In Proceedings of First Joint Conference on Lexical and Computational Semantics (*SEM) (pp. 114-123). Association for Computational Linguistics.

Murphy, B., Talukdar, P. P., & Mitchell, T. M. (2012b). Learning Effective and Interpretable Semantic Models using Non-Negative Sparse Embedding. In COLING (pp. 1933-1950).

Murphy, G. (2004). The big book of concepts (p. 555). MIT Press.

Nevins, A., Pesetsky, D., & Rodrigues, C. (2009). Pirahã exceptionality: A reassessment. Language, 85(2), 355-404.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium.

Pliatsikas, C., & Luk, G. (2016). Executive control in bilinguals: A concise review on fMRI studies. Bilingualism: Language and Cognition, 1-7.

Pulvermüller, F. (2005). Brain mechanisms linking language and action. Nature Reviews Neuroscience, 6, 576–582.

Raizada, R. D., & Connolly, A. C. (2012). What makes different people's representations alike: neural similarity space solves the problem of across-subject fMRI decoding. Journal of cognitive neuroscience, 24(4), 868-877.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. Proceedings of the National Academy of Sciences of the United States of America, 104, 1436–1441.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., ... & Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage, 15(1), 273-289.

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., & Mitchell, T. (2014a). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. PloS one, 9(11), e112575.

Wehbe, L., Vaswani, A., Knight, K., & Mitchell, T. (2014b). Aligning context-based statistical models of language with brain activity during reading. Proceedings of EMNLP.

Wilkins, J. (1668). An essay towards a real character, and a philosophical language. English linguistics, 1500-1800; a collection of facsimile reprints; no. 119, 1968, Scholar Press.

Zinszer, B. D., Anderson, A. J., Kang, O., Wheatley, T., & Raizada, R. D. (2016). Semantic Structural Alignment of Neural Representational Spaces Enables Translation between English and Chinese Words. Journal of Cognitive Neuroscience. Advance Online Publication. DOI: 10.1162/jocn_a_01000.

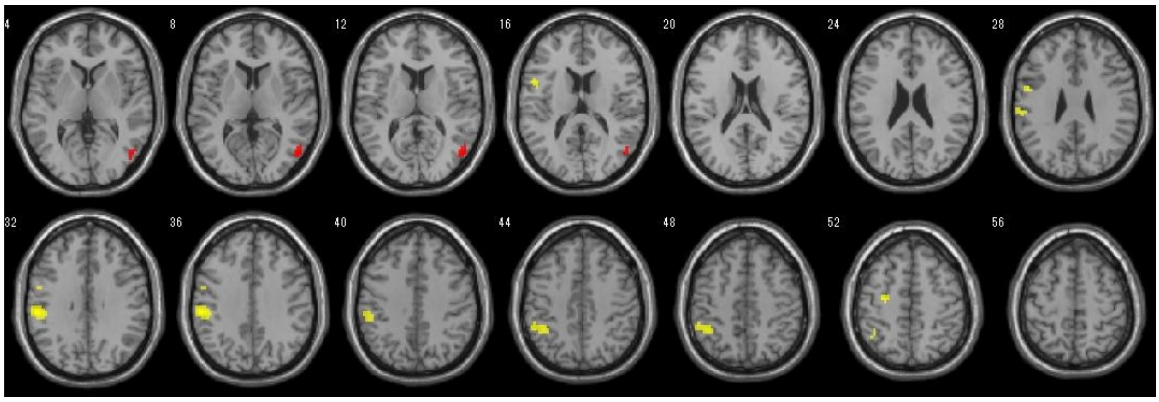

Figure 1: Result maps of two contrasts (red : mammal > tool ; yellow : tool > mammal) computed by using a random effect analysis in SPM8 with p < 0.001 uncorrected from 10 participant sessions. Mammal-specific activity is seen in inferior temporal cortex, and tool-specific activity is seen in motor areas

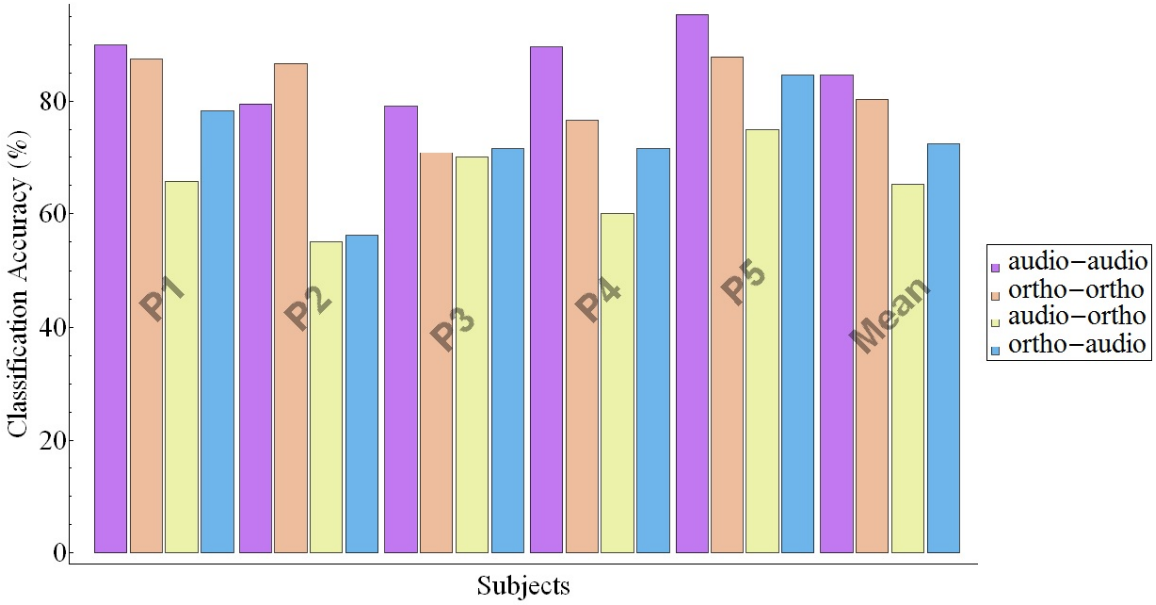

Figure 2: mammal/tool classification accuracy for within-participant analyses from Experiment I, using whole-trial boxcar (Figure remade based on Akama et al 2012).

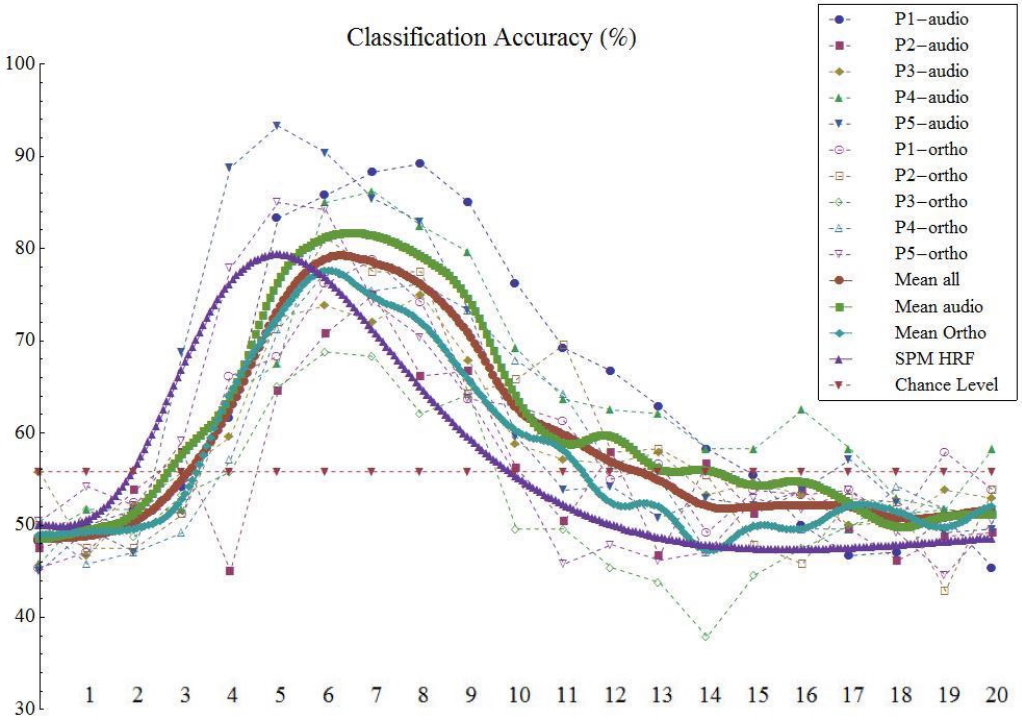

Figure 3: Classification accuracy for within-session analyses from Experiment I, using single-volume sliding window (Figure remade based on Akama et al 2012).

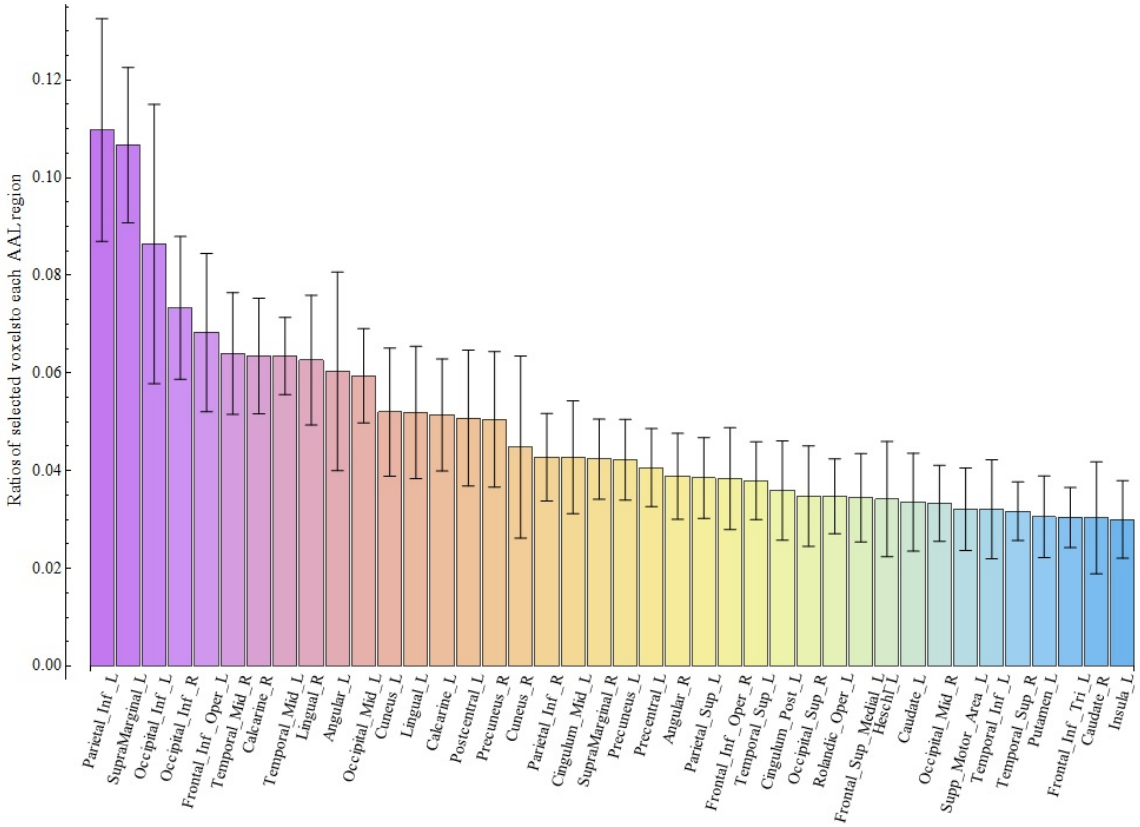

Figure 4: The 40 top mean ratios of these informative voxels to all the same-sized voxels

attributed to each anatomical area in AAL brain atlas. Error bars represent standard error (data from Experiment I)

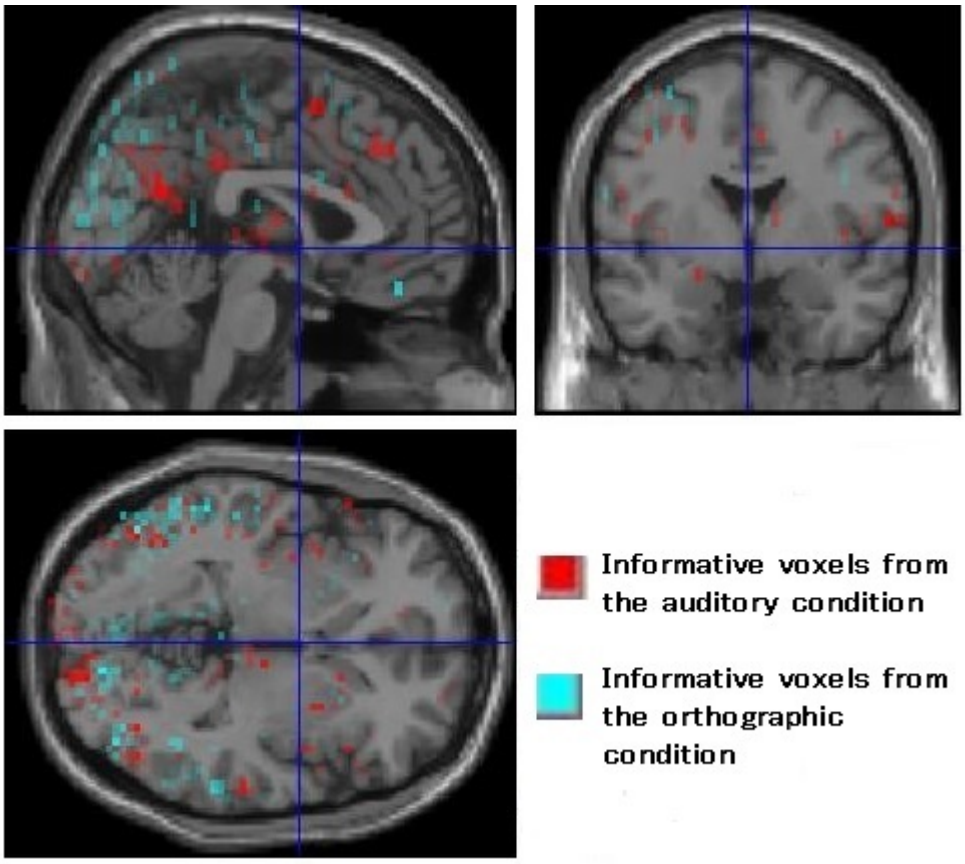

Figure 5: Modality-specific distribution maps of the 1000 most informative voxels, Experiment I.

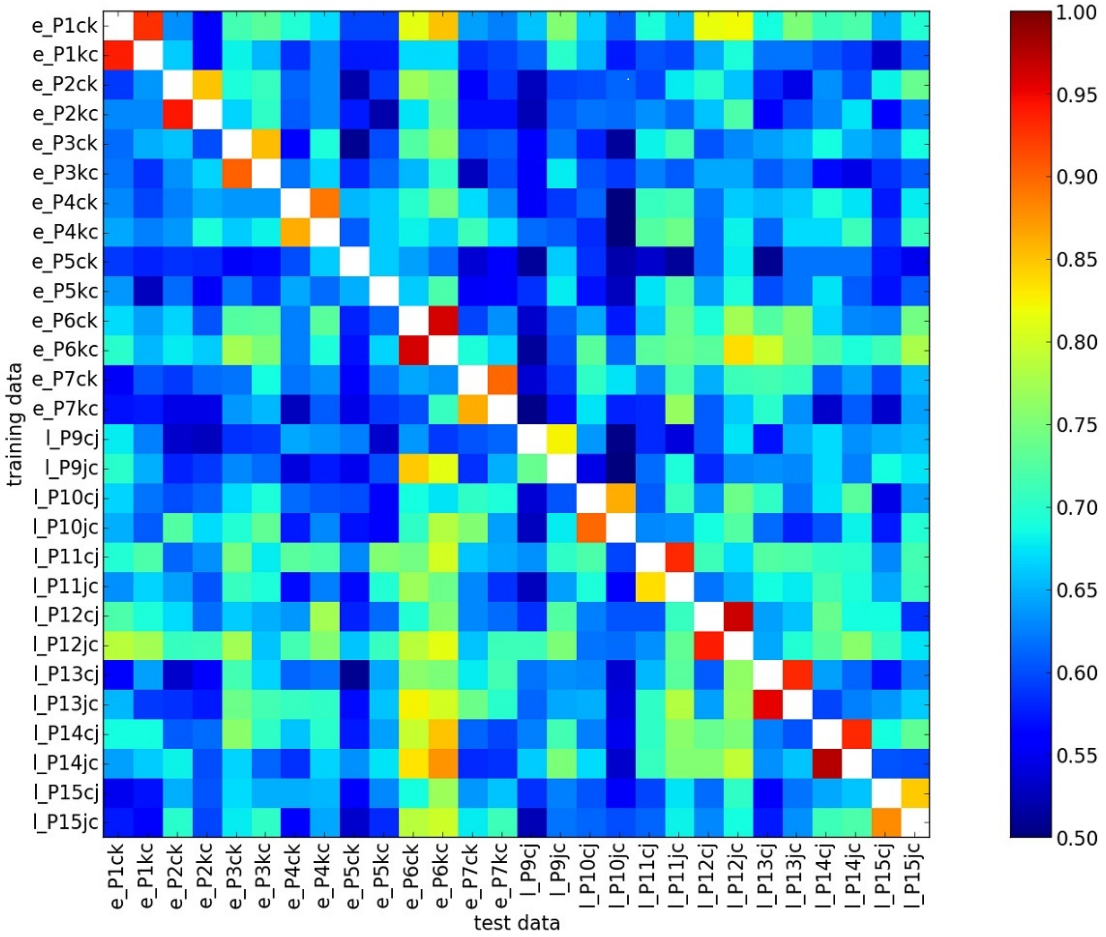

Figure 6: Cross-session classification accuracies between participants comprising early Korean-Chinese bilinguals ('e', Experiment IIa), and late Chinese-Japanese bilinguals ('l', Experiment IIb). The diagonal (within-session analysis) is not populated. The off-diagonals show the higher accuracy seen for within-participant analyses. (Replicated from Akama et al., 2014).

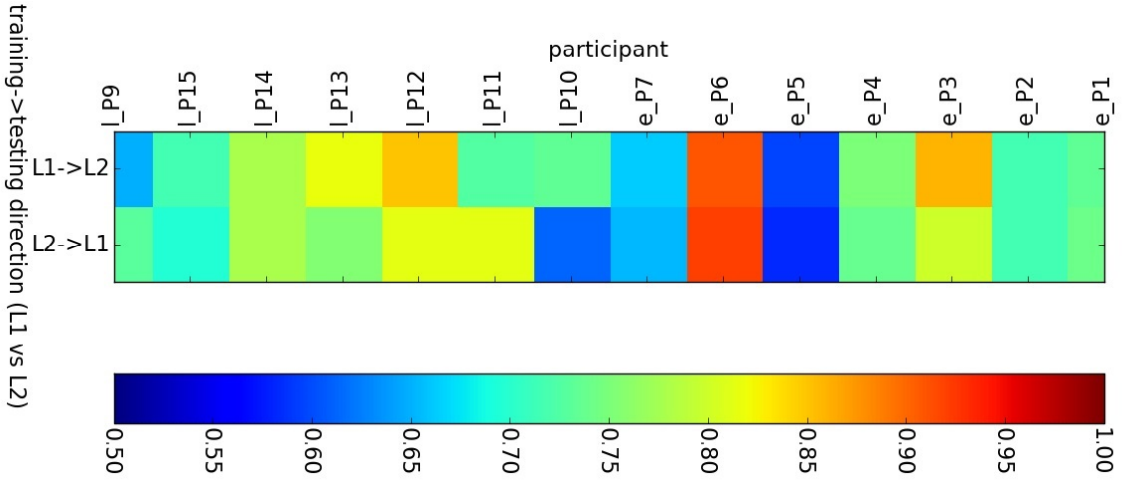

Figure 7: Groupwise classification accuracies, training on a set of Korean-Chinese bilinguals ('e', Experiment IIa), and late Chinese-Japanese bilinguals ('l', Experiment IIb),

and testing on a single left-out participant. (Replicated from Akama et al., 2014).

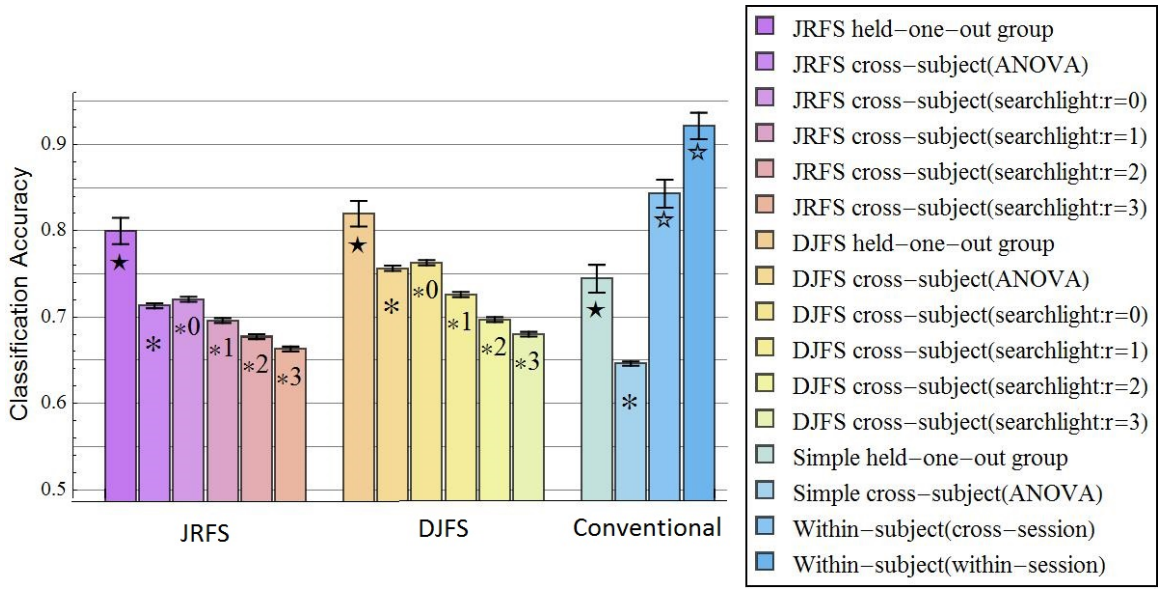

Figure 8: Summary of average classification accuracies for a range of analysis settings in Experiments IIa, IIb. JRFS and DJFS refer to feature selection strategies introduced in the next section. (Replicated from Akama et al., 2014).

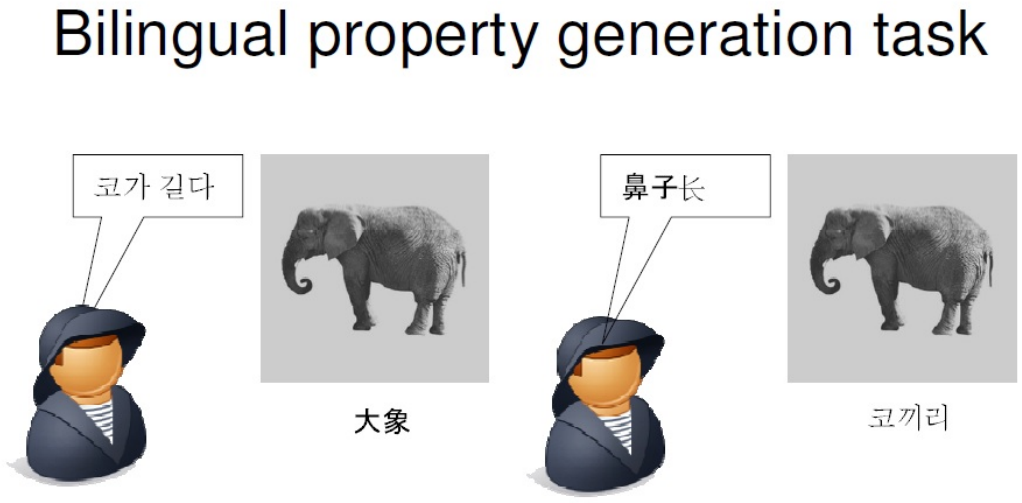

Figure 9: ‘Language A → Language B’ indicates that the participant was presented with stimuli captions in language A and was asked to perform covert property generation in language B.

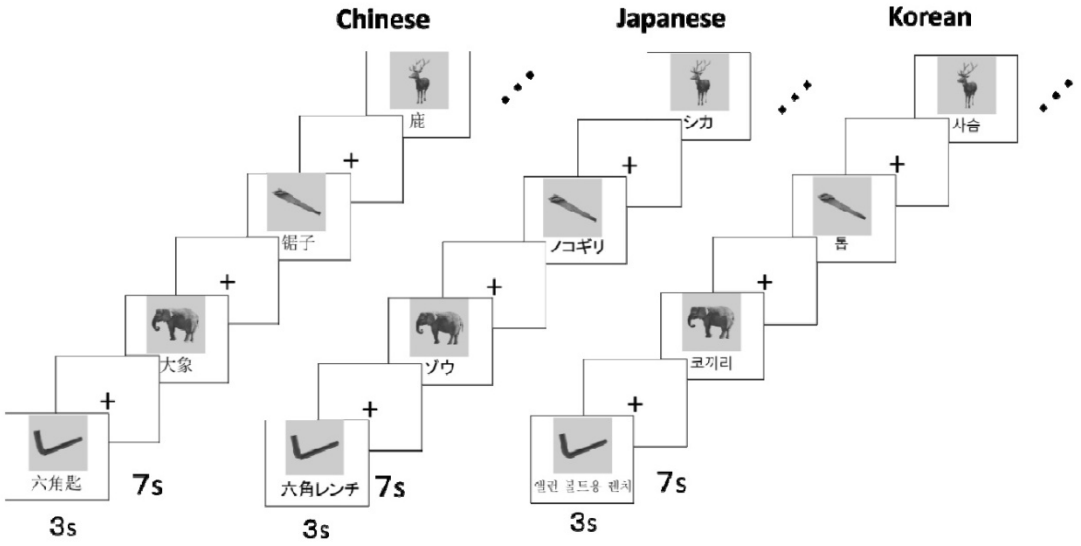

Figure 10: Trial structure and stimulus format for experiments IIa and IIb.